THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學



# Disentangled Speaker Embedding for Robust Speaker Verification

*Lu YI, Man-Wai MAK*
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University

## Introduction

- Speaker verification (SV) is a kind of biometric authentication that uses one's voice to verify a claimed speaker's identity.
- Domain mismatch (caused by, e.g., different microphone types) would degrade the performance of SV systems.

## Motivations

**Limitations** of some state-of-the-art domain adaptation methods:

- Only focus on common feature space without considering the domain-specific components;
- Ignore the difference in label distributions.

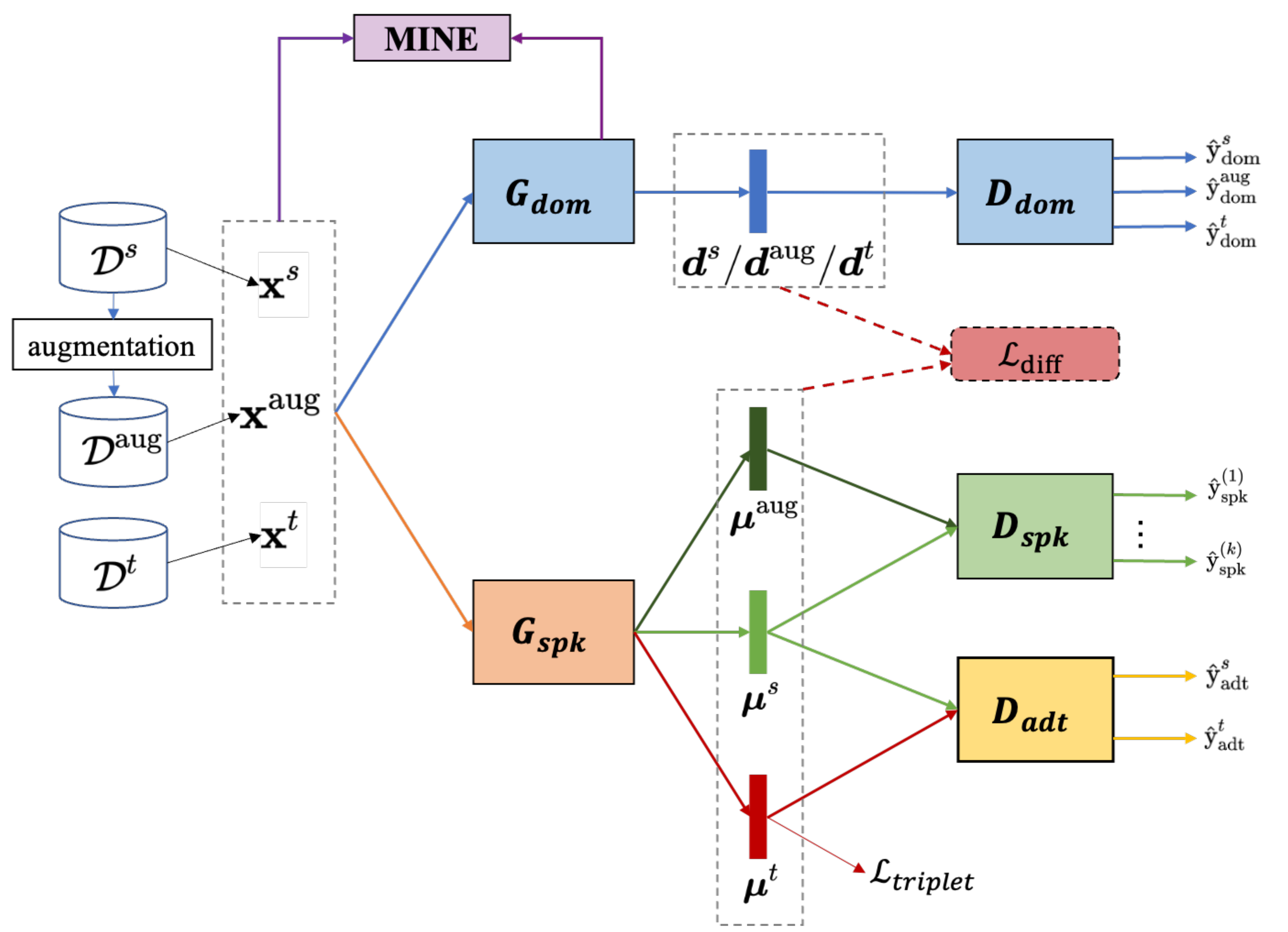
**Objectives** of this study:

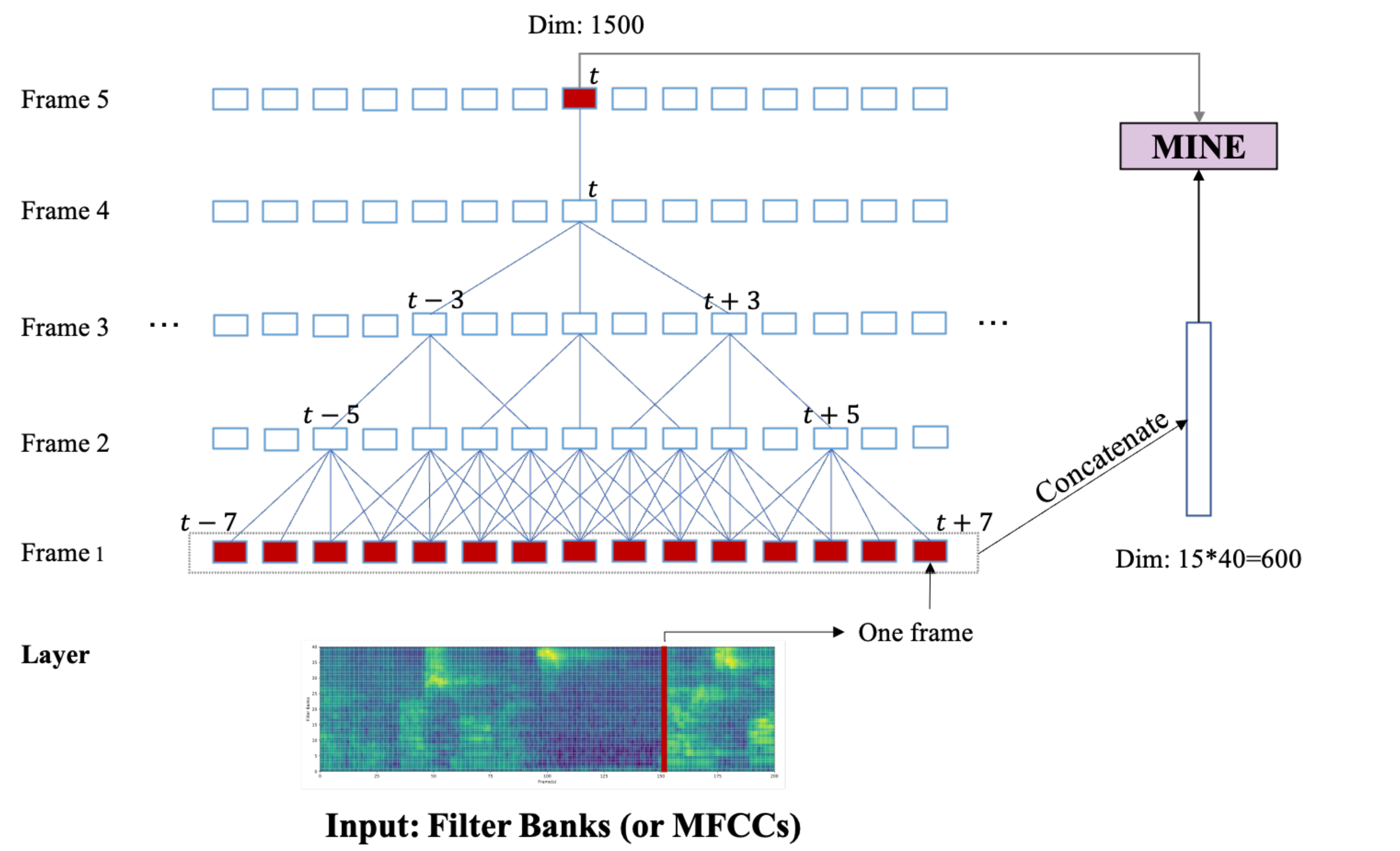
- Propose a novel framework to disentangle domain-invariant speaker features and domain-specific features;
- Incorporate domain adaptation directly into the training of speaker embedding extractor;
- Apply self-supervised learning to overcome the label mismatch problem without using labels from the target domain;
- Introduce a frame-based mutual information neural estimator that maximize the mutual information between the frame-level features and input acoustic features to learn informative features.


### References

1. Jonathan Huang and Tobias Bocklet, "Intel far-field speaker recognition system for VOiCES challenge 2019," in *Proc. Interspeech*, 2019, pp. 2473–2477.
2. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.


## InfoMax Domain Separation and Adaptation Network




*Frame-based mutual information neural estimation*

$$\min_{G_{\text{spk}}}\max_{D_{\text{adt}}}\mathcal{L}_{\text{adt}} = \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}[\log D_{\text{adt}}(G_{\text{spk}}(\mathbf{x}^s))] + \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[\log\left[1 - D_{\text{adt}}\left(G_{\text{spk}}\left(\mathbf{x}^t\right)\right)\right]\right]$$

$$\mathcal{L}_{\text{spk}} = \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[-\sum_{k=1}^{K} y_{\text{spk}}^{(k)}\log D_{\text{spk}}(G_{\text{spk}}(\mathbf{x}^s))_k\right] + \mathbb{E}_{\mathbf{x}^{\text{aug}}\sim\mathcal{D}^{\text{aug}}}\left[-\sum_{k=1}^{K} y_{\text{spk}}^{(k)}\log D_{\text{spk}}(G_{\text{spk}}(\mathbf{x}^{\text{aug}}))_k\right]$$

$$\mathcal{L}_{\text{dom}} = \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^t\right)\right)_0\right] + \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}[-\log D_{\text{dom}}(G_{\text{dom}}(\mathbf{x}^s))_1] + \mathbb{E}_{\mathbf{x}^{\text{aug}}\sim\mathcal{D}^{\text{aug}}}[-\log D_{\text{dom}}(G_{\text{dom}}(\mathbf{x}^{\text{aug}}))_2]$$

$$\mathcal{L}_{\text{MINE}} = -I_\Theta^{\text{JS}}(X,Z), \quad I_\Theta^{\text{JS}}(X,Z) = \mathbb{E}_{\mathbb{P}_{XZ}}[-\text{sp}(-T_\theta)] - \mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Z}[\text{sp}(T_\theta)]$$

$$\mathcal{L}_{\text{triplet}} = \max(d(a,p) - d(a,n) + \text{margin}, 0)$$

$$\mathcal{L}_{\text{diff}}^{\text{ort}} = \sum_i\left|(\boldsymbol{\mu}_i^s)^\top\mathbf{d}_i^s\right| + \sum_j\left|\left(\boldsymbol{\mu}_j^{\text{aug}}\right)^\top\mathbf{d}_j^{\text{aug}}\right| + \sum_k\left|(\boldsymbol{\mu}_k^t)^\top\mathbf{d}_k^t\right|; \quad \mathcal{L}_{\text{diff}}^{\text{MI}} = -I_\Theta^{\text{JS}}(\boldsymbol{\mu}^s,\mathbf{d}^s) - I_\Theta^{\text{JS}}(\boldsymbol{\mu}^{\text{aug}},\mathbf{d}^{\text{aug}}) - I_\Theta^{\text{JS}}(\boldsymbol{\mu}^t,\mathbf{d}^t)$$

## Experimental Setup

**Source domain data $\mathcal{D}^s$**

- Voxceleb1 dev, Voxceleb2 dev & test set.

**Augmented source domain data $\mathcal{D}^{aug}$**

- by adding noise, babble, and music from MUSAN [20] and reverberation from the RIR dataset to speech in $\mathcal{D}^s$.

**Target domain data $\mathcal{D}^t$**

- VOiCES dev set.

**Input acoustic features**

- 40-dimensional filter bank features with a frame length of 25ms at 10ms shift;
- using Kaldi energy-based voice activity detection (VAD) to remove silence frames;
- using small chunks of acoustic sequences with a chunk length of 200 frames for training.

## Results and Discussions

| Row | System | Source domain | Target domain | MINE | $\mathcal{L}_{\text{triplet}}$ | $D_{\text{adt}}$ | $\mathcal{L}_{\text{diff}}$ | VOiCES dev. EER(%) | VOiCES dev. minDCF | VOiCES eval. EER(%) | VOiCES eval. minDCF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TDNN [1] | $\mathcal{D}^s$ | $\mathcal{D}^{\text{aug}}$ | × | × | ✓ | × | 3.18 | - | 7.15 | - |
| 2 | TDNN ($G_{\text{spk}}$) | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | × | × | × | × | × | 2.35 | 0.2596 | 6.42 | 0.4398 |
| 3 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | × | ✓ | × | × | 2.16 | 0.2627 | 6.32 | 0.4305 |
| 4 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | × | ✓ | × | × | × | 2.29 | 0.2453 | 5.98 | 0.4351 |
| 5 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | × | ✓ | ✓ | × | 2.31 | 0.2645 | 6.29 | 0.4399 |
| 6 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | ✓ | ✓ | ✓ | × | 2.54 | 0.2671 | 6.23 | 0.4379 |
| 7 | InfoMax–DSAN ($G_{\text{dom}}$&$G_{\text{spk}}$) | $\mathcal{D}^s$ | $\mathcal{D}^{\text{aug}}$ | ✓ | × | ✓ | MI | 2.17 | 0.2418 | 5.93 | 0.4265 |
| 8 | | $\mathcal{D}^s$ | $\mathcal{D}^{\text{aug}}$ | ✓ | × | ✓ | ort | 2.33 | 0.2577 | 5.98 | 0.4131 |
| 9 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | ✓ | × | ✓ | MI | 3.29 | 0.3278 | 6.75 | 0.4614 |
| 10 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | × | ✓ | ✓ | MI | 2.31 | 0.2490 | 6.02 | 0.4161 |
| 11 | | $\{\mathcal{D}^s,\mathcal{D}^{\text{aug}}\}$ | $\mathcal{D}^t$ | ✓ | ✓ | ✓ | MI | **2.06** | **0.2375** | **5.69** | **0.4127** |

- The proposed frame-based MINE can effectively help extract informative features. It can either help extract informative speaker embeddings, or help disentangle redundant features from speaker features;
- Ignoring the label mismatch problem would degrade performance. Applying self-supervised learning can help address this problem;
- The domain adaptation system that considers the domain-specific features performs better than the system only focus on common feature space.